GS01 0163
Analysis of Microarray Data

Keith Baggerly and Brad Broom
Department of Bioinformatics and Computational Biology
UT M. D. Anderson Cancer Center
kabagg@mdanderson.org
bmbroom@mdanderson.org

11 November 2010
Lecture 21: Batch Effects

- What are batch effects? How do we plan for them?
- How do we detect them?
- How do we correct for them?
- DWD – Benito et al, Bioinf 20:105-14, 2004
- Review – Leek et al, Nat Rev Gen 11:733-9, 2010
- TCGA batches and genes
Batch Effect Sources and Planning

This is a free-for-all slide, in that we’ve talked about these several times now.

Some causes?

Some ways of planning?

An ovarian cancer case study:
Dressman et al, JCO 2007
Signatures and Pathways for Response in Ovarian Cancer

An Integrated Genomic-Based Approach to Individualized Treatment of Patients With Advanced-Stage Ovarian Cancer

Holly K. Dressman, Andrew Berchuck, Gina Chan, Jun Zhai, Andrea Bild, Robyn Sayer, Janiel Cragun, Jennifer Clarke, Regina S. Whitaker, LiHua Li, Jonathan Gray, Jeffrey Marks, Geoffrey S. Ginsburg, Anil Potti, Mike West, Joseph R. Nevins, and Johnathan M. Lancaster


Using profiles of 119 ovarian tumors, they looked for signatures of response to cisplatin-based chemo. The also looked at the deregulation levels of 5 pathways (Src, β-catenin, Myc, E2F3, and Ras), trying to relate them to survival.
Checking Agreement

Ours vs theirs. We expected better (fewer outliers).
Looking at Their Other Quants

Finding the Best Match, CEL RMA Column 1

Which one would you pick?
Looking at The “Best” Fit

Two RMA Quantifications: 872 From CEL, 2476 From XLS

Same array. *Different* names (2476 from XLS, 872 from CEL).
Grab One More Thing: Run Date

Should we be worried?
Some Berchuck et al Clinical Info

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2392</td>
<td>Early Stage</td>
</tr>
<tr>
<td>2393</td>
<td>Early Stage</td>
</tr>
<tr>
<td>1772</td>
<td>Long</td>
</tr>
<tr>
<td>1773</td>
<td>Long</td>
</tr>
<tr>
<td>1774</td>
<td>Long</td>
</tr>
<tr>
<td>1775</td>
<td>Long</td>
</tr>
<tr>
<td>1776</td>
<td>Long</td>
</tr>
<tr>
<td>1777</td>
<td>Long</td>
</tr>
<tr>
<td>1778</td>
<td>Long</td>
</tr>
<tr>
<td>1779</td>
<td>Long</td>
</tr>
<tr>
<td>1780</td>
<td>Long</td>
</tr>
<tr>
<td>1781</td>
<td>Long</td>
</tr>
<tr>
<td>1900</td>
<td>Long</td>
</tr>
</tbody>
</table>
The Berchuck et al Dat Headers

0074_1772_h133a_872.cel:DatHeader=[0..37764] .. 09/20/02 11
0074_1773_h133a_922.cel:DatHeader=[0..33251] .. 09/20/02 11
0074_1774_h133a_1451.cel:DatHeader=[0..43335] .. 09/20/02 11
0074_1775_h133a_1526.cel:DatHeader=[0..45012] .. 09/20/02 11
0074_1776_h133a_1784.cel:DatHeader=[0..46104] .. 09/20/02 11
0074_1777_h133a_1834.cel:DatHeader=[0..42469] .. 09/20/02 11
0074_1778_h133a_1846.cel:DatHeader=[0..36713] .. 09/20/02 11
0074_1779_h133a_2075.cel:DatHeader=[0..37459] .. 09/20/02 11
0074_1780_h133a_2204.cel:DatHeader=[0..43583] .. 09/20/02 11
0074_1781_h133a_2419.cel:DatHeader=[0..46101] .. 09/20/02 11
0074_1827_h133a_.08.cel:DatHeader=[0..46104] .. 10/23/02 12
0074_1828_h133a_860.cel:DatHeader=[0..46102] .. 10/23/02 12
The Berchuck et al Heatmap

Can you spot clusters? Expression tracks with batch.
Survival is confounded with date.
Is There Division by CR/NR?

Maybe something.
Is There Division by Date?

Erm, yep.
Division by CR/NR After Date?

Not really.
How Many Batches? (Controls)

Using Affy controls, we see 7 blocks.
How Many Batches? (Top ANOVAs)

The smallest ANOVA p-values show the same.
Is There Division by Batch?

We think we caught most of it.
Division by CR/NR After Batch?

Maybe some, but the FDR isn’t small.
Batches Affect Scores

Offsets are quite visible
Early Batch Effects Are Larger

Remember, batches are confounded with Survival!
So, how did we get Run Date?

> celDatHeaders <- celFiles
> for (i1 in 1:length(celDatHeaders)) {
    temp <- read.celfile.header(file.path("DukeWebSite", "PlatinumJCO",
    celFiles[i1]), info = "full")
    celDatHeaders[i1] <- temp$DatHeader
}
Eyeballing DatHeader Lines

> celDatHeaders[1]

872

"[0..37764] 0074_1772_H133A_872:CLS=4733 RWS=4733 XIN=3 YIN=3 VE=17 2.0 09/20/02 11:43:50 \024 \024 HG-U133A.1sq \024 \024 \024 \024 \024 \024 \024 \024 \024 6"

celDatHeaders[119]

M810

"[0..29187] Robyn 810 CR-2:CLS=5391 RWS=5391 XIN=2 YIN=2 VE=30 2.0 04/20/04 11:29:32 50101330 M10 \024 \024 HG-U133A.1sq \024 \024 \024 \024 \024 \024 \024 \024 \024 6"
Extracting Dates

> tempDate1 <- strsplit(celDatHeaders, "2\\.0 ")
> tempDate2 <- unlist(lapply(tempDate1, function(x) { x[2] } ))
> tempDate3 <- unlist(lapply(tempDate2, function(x) { substr(x, 1, 17) } ))
> tempDate3[1]

872
"09/20/02 11:43:50"
Formatting Dates

```r
> celRunDate <- as.Date(tempDate3, 
  format = "%m/%d/%y %H:%M:%S")
> names(celRunDate) <- names(tempDate3)
> celRunDate <- celRunDate[rownames(clinicalInfo)]
> celRunDate[1:3]
  0.08 860 872
"2002-10-23" "2002-10-23" "2002-09-20"
```
Tabulating Dates

> table(celRunDate)

celRunDate
10 9 9 1 3 11 10 1 16 6 5

Using Dates

```R
> names(celRunDate[celRunDate < "2004-03-09"])
[1] "0.08" "860" "872" "922" ...
[51] "3102" "3107" "3142" "3249"
> names(celRunDate[celRunDate >= "2004-03-09"])
[1] "D1805" "D1837" "D1859" "D2098" ...
[64] "M6199" "M810"
> sum(celRunDate < "2004-03-09")
[1] 54
> table(celRunDate, clinicalInfo$Response)
celRunDate CR NR
 2002-09-20 10  0
 2002-10-23  2  7
 2002-11-12  8  1
 2002-12-16  1  0
```
So, how did we Correct for Batch?

```r
> runBatch <- rep(6, 119)
> runBatch[celRunDate == "2002-09-20"] <- 1
> runBatch[celRunDate == "2002-10-23"] <- 2
> runBatch[celRunDate == "2002-11-12"] <- 2
> runBatch[celRunDate == "2002-12-16"] <- 3
> runBatch[celRunDate == "2002-12-21"] <- 3
> runBatch[celRunDate == "2003-01-03"] <- 3
> runBatch[celRunDate == "2003-05-30"] <- 4
> runBatch[celRunDate == "2003-06-26"] <- 4
> runBatch[celRunDate == "2004-03-09"] <- 5
> runBatch[celRunDate == "2004-03-16"] <- 5
> runBatch[celRunDate == "2004-04-20"] <- 5
> runBatch[celRunDate == "2004-05-18"] <- 6
> runBatch[celRunDate == "2004-05-21"] <- 6
```
> runBatch[celRunDate == "2004-05-27"] <- 6
> runBatch[celRunDate == "2004-06-22"] <- 6
> runBatch[celRunDate == "2004-06-23"] <- 6
> runBatch <- as.factor(runBatch)
> names(runBatch) <- names(celRunDate)
Modeling Residuals

Fit gene expression values as a function of runBatch, then fit residuals after correcting for batch.

> batchModelForm <- Y ~ runBatch
> batchModelLMAAll <-
  MultiLinearModel(batchModelForm,
  data.frame(runBatch = runBatch), ovcaRMAFromCEL)
> ovcaRMAFromCELResids <- ovcaRMAFromCEL -
  t(batchModelLMAAll@predictions)
> responseModelForm <- Y ~ Response
> responseModelLMAAll <-
  MultiLinearModel(responseModelForm,
  data.frame(Response = clinicalInfo$Response),
  ovcaRMAFromCELResids)
Other Approaches

Principal Components/Singular Value Decomposition

Excellent for getting a high-level view of the structure in your data, and seeing if there are outliers or outlying clusters.

Hence, useful for revealing batches, but not so useful for removing them – it doesn’t use labels at all.
Where SVD Breaks

Benito et al, Bioinf 2004, fig 1. The batch effect isn’t clearly dominant. LDA is ok here.
Where LDA Breaks

Benito et al, Bioinf 2004, fig 2. Low-d structure in high-d data. LDA sidetracked.
Distance Weighted Discrimination (DWD)

Similar to a Support Vector Machine (SVM) approach, but not solely driven by extrema (maximizing the minimum distance). Instead, maximize the sum of the inverse distances — lets all points contribute, but still stays robust to outliers.

Matlab implementation code exists, involves some sophisticated optimization algorithms.

Focuses on two groups, works better if groups are moderately large.
Benito et al, Bioinf 2004, fig 6a. Datasets from two platforms (logratio), each median centered.
Real Data post DWD

An Alternative – COMBAT

Starts with an adjustment similar to what we tried in the first place with the linear models, with two shifts.

First, it allows for shifts in scale as well as center.

Second, it borrows strength across genes and shrinks towards the central location and scale adjustments.
A COMBATive Triptych

Bad, Good, Good
A Shrinking Black Hole

Which values are “better”? Why?
The Value of Shrinkage

Borrowing can let you recognize (and downweight) outliers.
Implementing COMBAT

This is actually pretty easy, since a set of R scripts is available from Evan Johnson’s web site at BYU:

http://statistics.byu.edu/johnson/ComBat
Review: Leek et al

How pervasive are batch effects?

How big are they?

How can we fix them?

When can we fix them?
Will Normalization Fix Things?

Leek et al Fig 1: Batches can survive normalization.
Are New Assays Immune?

Figure 2 | Batch effects for second-generation sequencing data from the 1000 Genomes Project. Each row is a different HapMap sample processed in the same facility with the same platform. See Supplementary information S1 (box) for a description of the data represented here. The

Leek et al Fig 2: Next gen — sd +/- in orange/blue.
Are Pathways Immune?

Figure 3 | **Batch effects also change the correlations between genes.** We normalized every gene

Leek et al Fig 3: Correlations can change sign
What Was Examined?

<table>
<thead>
<tr>
<th>Study description*</th>
<th>Known variable used as a surrogate</th>
<th>Principal components used as a surrogate</th>
<th>Association with outcome</th>
<th>Refs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Surrogate (%)</td>
<td>Principal components rank of surrogate (correlation) (%)</td>
<td>Significant features (%)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Confounding (%)</td>
<td>Principal components rank of outcome (correlation) (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Susceptible features (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data set 1: gene expression microarray, Affymetrix (N_p = 22,283)</td>
<td>Date</td>
<td>29.7</td>
<td>50.5</td>
<td>1 (0.570)</td>
</tr>
<tr>
<td>Data set 2: gene expression, Affymetrix (N_p = 4167)</td>
<td>Date</td>
<td>77.6</td>
<td>73.7</td>
<td>1 (0.922)</td>
</tr>
<tr>
<td>Data set 3: mass spectrometry (N_p = 15,154)</td>
<td>Processing group</td>
<td>100</td>
<td>51.7</td>
<td>2 (0.344)</td>
</tr>
<tr>
<td>Data set 4: copy number variation, Affymetrix (N_p = 945,808)</td>
<td>Date</td>
<td>29.2</td>
<td>99.5</td>
<td>2 (0.921)</td>
</tr>
<tr>
<td>Data set 5: copy number variation, Affymetrix (N_p = 945,808)</td>
<td>Date</td>
<td>12.2</td>
<td>83.8</td>
<td>2 (0.553)</td>
</tr>
<tr>
<td>Data set 6: gene expression, Affymetrix (N_p = 22,277)</td>
<td>Processing group</td>
<td>NA</td>
<td>83.8</td>
<td>5 (0.369)</td>
</tr>
<tr>
<td>Data set 7: gene expression, Agilent (N_p = 17,594)</td>
<td>Date</td>
<td>NA</td>
<td>62.8</td>
<td>2 (0.248)</td>
</tr>
<tr>
<td>Data set 8: DNA methylation, Agilent (N_p = 27,578)</td>
<td>Processing group</td>
<td>NA</td>
<td>78.6</td>
<td>3 (0.381)</td>
</tr>
<tr>
<td>Data set 9: DNA sequencing, Solexa (N_p = 2,886)</td>
<td>Date</td>
<td>24.2</td>
<td>32.1</td>
<td>2 (0.846)</td>
</tr>
</tbody>
</table>

Leek et al Table 1: Batches are everywhere
How Do We Fix Things?

Design
Linear Models
ComBat
SVA
Which do I Prefer?

I like them all, but at present I lean towards either linear model approaches or ComBat.

This is for reasons of relative simplicity and ease of implementation.

In general, I think the choice of a particular method of batch adjustment is an order of magnitude less important than recognizing that batches may be present in the first place, and coming up with reasonable ideas as to what they are.
TCGA Batches

The TCGA samples are processed in batches.

This is largely by necessity, because not all of the samples are coming into the processing center at once, and they don’t want to wait for all 500 before starting.

Batches typically involve about 30 samples.

The batches have not been assembled with balance in mind.

For some batches, mild correction will likely be necessary.
TCGA Genes

“GBM Phase I + II.xls” (from the TCGA web site)

<table>
<thead>
<tr>
<th>Phase</th>
<th>Name</th>
<th>Chr</th>
<th>Start</th>
<th>Stop</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phase1</td>
<td>ABI1</td>
<td>chr10</td>
<td>27077005</td>
<td>27190298</td>
</tr>
<tr>
<td>Phase1</td>
<td>ABL1</td>
<td>chr9</td>
<td>132579028</td>
<td>132751391</td>
</tr>
<tr>
<td>Phase1</td>
<td>ADAM15</td>
<td>chr1</td>
<td>153289987</td>
<td>153302102</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>locusLinkId</th>
<th>omimId</th>
<th>proteinAcc</th>
</tr>
</thead>
<tbody>
<tr>
<td>10006</td>
<td>603050</td>
<td>NP_005461</td>
</tr>
<tr>
<td>25</td>
<td>189980</td>
<td>NP_009297</td>
</tr>
<tr>
<td>8751</td>
<td>605548</td>
<td>NP_003806</td>
</tr>
</tbody>
</table>
TCGA Positions?

Shouldn’t they line up?